Pythagore F.D.

Apprendre à apprendre

CB018

Programmation R pour Hadoop

Durée: 3 jours

1660 €	27 au 29 mars	16 au 18 octobre
	3 au 5 juillet	18 au 20 décembre

Public:

Chefs de projet, data scientists, statisticiens, développeurs souhaitant comprendre les apports de R pour l'analyse des données, et savoir l'intégrer à un environnement Hadoop.

Objectifs:

Connaître les principales fonctions statistiques de R, et savoir utiliser des programmes R dans un environnement Hadoop, en s'appuyant sur le système distribué hdfs et le stockage avec HBase..

Connaissances préalables nécessaires:

Connaissance des bases Hadoop, et notions de calculs statistiques

Programme:

Présentation R	:	Le projet R Programming Calculs statistiques et génération de graphiques Points forts de R Programming Besoins du BigData Positionnement R programming par rapport à Hadoop
Mise en oeuvre de R	:	Travaux pratiques : installation et tests sur une plate-forme CentOS Utilisation de R en mode commande. Commandes de base. Syntaxe. Manipulations de nombres,vecteurs,tableaux,matrices.listes,etc ..
Intégration Hadoop	:	Association de la puissance du calcul distribué fourni par les outils hadoop, et de la richesse des outils d'analyse statistique de R. Différents moyens d'intégration : RHive : fonctions R de calculs statistiques s'appuyant sur HiveQL RHadoop : packages rmr2, rhdfs pour utiliser le système distribué hdfs depuis R, rhbase pour accéder à HBase depuis les programmes en R.
Travaux pratiques avec Hadoop	:	Installation d'un cluster, rmr2:traduction programmes R en mapreduce, rhdfs:API d'accès R à des données stockéss sur HDFS rhbase:API d'accès à des données stockées sur HBase
Evolutions	:	Les acteurs : IBM avec BigInsights, Revolution R avec ScaleR